

On the relation of nonanticipative rate distortion function and filtering theory

Charalambos D. Charalambous¹ and Photios A. Stavrou²

Abstract—In this paper the relation between nonanticipative rate distortion function (RDF) and Bayesian filtering theory is investigated using the topology of weak convergence of probability measures on Polish spaces. The relation is established via an optimization on the space of conditional distributions of the so-called directed information subject to fidelity constraints. Existence of the optimal reconstruction distribution of the nonanticipative RDF is shown, while the optimal nonanticipative reproduction conditional distribution for stationary processes is derived in closed form. The realization procedure of nonanticipative RDF is described, while an example is introduced to illustrate the concepts.

I. INTRODUCTION

This paper is concerned with the abstract formulation of nonanticipative rate distortion function (RDF) on Polish spaces (complete separable metric spaces) and its relation to filtering theory. In the past, rate distortion (or distortion rate) functions and filtering theory have evolved independently. Specifically, classical RDF addresses the problem of reconstruction of a process subject to a fidelity criterion without much emphasis on the realization of the reconstruction conditional distribution via nonanticipative operations. On the other hand, filtering theory is developed by imposing real-time realizability on estimators with respect to measurement data. Specifically, least-squares filtering theory deals with the characterization of the conditional distribution of the unobserved process given the measurement data, via a stochastic differential equation which depends on the observation data [1] via nonanticipative operations.

Although, both reliable communication and filtering (state estimation for control) are concerned with the reconstruction of processes, the main underlying assumptions characterizing them are different.

Historically, the work of R. Bucy [2] appears to be the first to consider the direct relation between distortion rate function and filtering, by carrying out the computation of a realizable distortion rate function with square criteria for two samples of the Ornstein-Uhlenbeck process. The work of A. K. Gorbunov and M. S. Pinsker [3] on ϵ -entropy defined via a nonanticipative constraint on the reproduction distribution of the RDF, although not directly related to the realizability question pursued by Bucy, computes the

nonanticipative RDF for stationary Gaussian processes via power spectral densities.

The objective of this paper is to investigate the connection between nonanticipative RDF and filtering theory for general distortion functions and random processes on abstract Polish spaces using the topology of weak convergence. The connection is established via optimization of directed information [4] over the space of conditional distributions which satisfy an average distortion constraint.

The main results discussed in this paper are the following.

- (1) Existence of optimal reconstruction distribution minimizing directed information using the topology of weak convergence of probability measures on Polish spaces;
- (2) Closed form expression of the optimal reconstruction conditional distribution for stationary processes;
- (3) Realization procedure of the filter;
- (4) Example to demonstrate the realization of the filter.

Motivation. This work is motivated by applications in which estimators are desired to have specific accuracy, such as processing information from sensor networks [5], and by control over limited rate communication channel applications [6], [7]. It is important to note that over the years several papers have appeared in the literature utilizing information theoretic measures for estimator and control applications [8], [9]. First, we give a brief high level discussion on nonanticipative RDF and filtering theory, and discuss their connection.

Consider a discrete-time process $X^n \triangleq \{X_0, X_1, \dots, X_n\} \in \mathcal{X}_{0,n} \triangleq \times_{i=0}^n \mathcal{X}_i$, and its reconstruction $Y^n \triangleq \{Y_0, Y_1, \dots, Y_n\} \in \mathcal{Y}_{0,n} \triangleq \times_{i=0}^n \mathcal{Y}_i$ where \mathcal{X}_i and \mathcal{Y}_i are Polish spaces.

Bayesian Estimation Theory. In classical filtering, one is given a mathematical model that generates the process X^n , $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \dots, n\}$, often induced via discrete-time recursive dynamics, a mathematical model that generates observed data obtained from sensors, say, Z^n , $\{P_{Z_i|Z^{i-1}, X^i}(dz_i|z^{i-1}, x^i) : i = 0, 1, \dots, n\}$, while Y^n are the causal estimates of some function of the process X^n based on the observed data Z^n . The classical Kalman Filter is a well-known example, where $\hat{X}_i = \mathbb{E}[X_i|Z^{i-1}]$, $i = 0, 1, \dots, n$, is the conditional mean which minimizes the average least-squares estimation error. Thus, in classical filtering theory both models which generate the unobserved and observed processes, X^n and Z^n , respectively, are given

*This work was financially supported by a medium size University of Cyprus grant entitled "DIMITRIS".

¹C. D. Charalambous is Professor with the Department of Electrical and Computer Engineering, University of Cyprus, 75 Kallipoleos Avenue, Nicosia, CYPRUS chadcha@ucy.ac.cy.

²P. A. Stavrou is PhD student with the Department of Electrical and Computer Engineering, University of Cyprus, 75 Kallipoleos Avenue, Nicosia, CYPRUS stavrou.fotios@ucy.ac.cy.

á priori. Fig. 1 is the block diagram of the filtering problem.

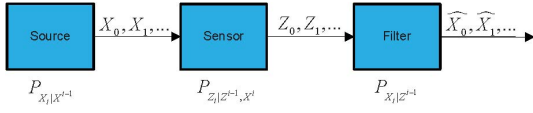


Fig. 1. Filtering problem.

Nonanticipative Rate Distortion Theory and Estimation. In nonanticipative rate distortion theory one is given a distribution for the process X^n , which induces $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \dots, n\}$, and determines the nonanticipative reconstruction conditional distribution $\{P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \dots, n\}$ which minimizes the directed information from X^n to Y^n subject to distortion or fidelity constraint. The filter $\{Y_i : i = 0, 1, \dots, n\}$ of $\{X_i : i = 0, 1, \dots, n\}$ is found by realizing the optimal reconstruction distribution $\{P_{Y_i|X^{i-1}, X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \dots, n\}$ via a cascade of sub-systems as shown in Fig. 2. Thus, in nonanticipative rate distortion theory the observation or mapping from $\{X_i : i = 0, 1, \dots, n\}$ to $\{Z_i : i = 0, 1, \dots, n\}$ is part of the realization procedure, while in filtering theory, this mapping is given á priori. Indeed, this is the main difference between Bayesian estimation theory and nonanticipative RDF for the purpose of estimation.

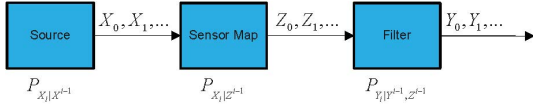


Fig. 2. Filtering via nonanticipative rate distortion function.

The precise problem formulation necessitates the definitions of distortion function or fidelity, and directed information. The distortion function or fidelity between x^n and its reconstruction y^n , is a measurable function defined by

$$d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \rightarrow [0, \infty], \quad d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho_{0,i}(x^i, y^i).$$

The directed information between X^n and Y^n , for a given distribution $P_{X^n}(dx^n)$, and conditional distribution $P_{Y^n|X^n}(dy^n|x^n)$, is defined by [10]⁴

$$\begin{aligned} I(X^n \rightarrow Y^n) &\triangleq \sum_{i=0}^n I(X^i; Y_i | Y^{i-1}) \\ &= \sum_{i=0}^n \int \log \left(\frac{P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i)}{P_{Y_i|Y^{i-1}}(dy_i|y^{i-1})} \right) P_{X^i, Y^i}(dx^i, dy^i) \\ &\equiv \mathbb{I}_{X^n \rightarrow Y^n}(P_{X_i|X^{i-1}, Y^{i-1}}, P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n). \end{aligned} \quad (1)$$

$$(2)$$

The notation $\mathbb{I}_{X^n \rightarrow Y^n}(\cdot, \cdot)$ illustrates the dependence of $I(X^n \rightarrow Y^n)$ on the two sequences

⁴Unless otherwise, integrals with respect to probability distributions are over the spaces on which these are defined.

of nonanticipative conditional distributions $\{P_{X_i|X^{i-1}, Y^{i-1}}(\cdot|\cdot, \cdot), P_{Y_i|Y^{i-1}, X^i}(\cdot|\cdot, \cdot) : i = 0, 1, \dots, n\}$. In information theory, directed information $\mathbb{I}_{X^n \rightarrow Y^n}(\cdot, \cdot)$ is often used as an information theoretic measure which describes the directivity of information flow via a sequence of channel outputs, defined by a nonanticipative sequence of feedback conditional distributions $P_{Y_i|Y^{i-1}, X^i}(\cdot|\cdot, \cdot)$ and feedforward conditional distributions $P_{X_i|X^{i-1}, Y^{i-1}}(\cdot|\cdot, \cdot)$, $i = 0, 1, \dots, n$. Directed information is also used in biological applications [11], [12] as a measure of causality, describing the cause and effect.

In this paper, it is assumed that $\forall i = 0, 1, \dots, n$

$$P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1}) = P_{x_i|x^{i-1}}(dx_i|x^{i-1}) - a.s. \quad (3)$$

The above assumption states that the process $\{X_i : i = 0, 1, \dots, n\}$ is conditionally independence of $Y^{i-1} = y^{i-1}$ given knowledge of $X^{i-1} = x^{i-1}$. Clearly, (3) is implied by the following conditional independence, $P_{Y_i|Y^{i-1}, X^\infty}(dy_i|y^{i-1}, x^\infty) = P_{y_i|y^{i-1}, x^i}(dy_i|y^{i-1}, x^i) - a.s., \forall i = 0, 1, \dots, n$. The last assumption implies that the reconstruction of Y_i does not depend on future values $X_{i+1}^\infty \triangleq \{X_{i+1}, X_{i+2}, \dots, X_\infty\}$, stating that Y_i is nonanticipative with respect to the process $\{Y_i : i = 0, 1, \dots, n\}$.

Given a probability distribution $P_{X^n}(dx^n)$ and a sequence of conditional distributions $\{P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n\}$ the directed information in the definition of nonanticipative RDF is given by

$$I(X^n \rightarrow Y^n) = \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, P_{Y_i|Y^{i-1}, X^i} : i = 0, 1, \dots, n). \quad (4)$$

Nonanticipative Rate Distortion Function. The nonanticipative RDF is defined by

$$R_{0,n}^c(D) \triangleq \inf_{P_{Y_i|Y^{i-1}, X^i} : i=0,1,\dots,n} \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, P_{Y_i|Y^{i-1}, X^i}) : \mathbb{E}\{d_{0,n}(X^n, Y^n) \leq D\} \quad (5)$$

The definition of the nonanticipative RDF is consistent with [3] in which non-anticipation is defined via the Markov chain $X_{n+1}^\infty \leftrightarrow X^n \leftrightarrow Y^n$, e.g., $P_{Y^n|X^\infty}(dy^n|x^\infty) = P_{Y^n|X^n}(dy^n|x^n)$. Therefore, by finding the solution of (5), then one can realize it via a channel from which one can construct an optimal filter via nonanticipative operations as in Fig. 2.

This paper is organized as follows. Section II discusses the formulation on abstract spaces. Section III establishes existence of optimal minimizing distribution, and Section IV derives the optimal minimizing distribution for stationary processes. Section V describes the realization of nonanticipative RDF, while Section VI provides an example. Lengthy proofs are omitted due to space limitations.

II. ABSTRACT FORMULATION

The source and reconstruction alphabets are sequences of Polish spaces [13] as defined in the previous section. Probability distributions on any measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$

are denoted by $\mathcal{M}_1(\mathcal{Z})$. For $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ measurable spaces, the set of conditional distributions $P_{Y|X}(\cdot|X=x)$ is denoted by $\mathcal{Q}(\mathcal{Y}; \mathcal{X})$ and these are equivalent to stochastic kernels on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ given $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

Given the process distributions $P_{X^n}(dx^n)$ and $\{P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \dots, n\}$ the following probability distributions are defined.

(P1): The reconstruction conditional probability distribution $\vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$:

$$\vec{P}_{Y^n|X^n}(dy^n|x^n) \triangleq \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) - a.s. \quad (6)$$

(P2): The joint probability distribution $P_{X^n, Y^n} \in \mathcal{M}_1(\mathcal{Y}_{0,n} \times \mathcal{X}_{0,n})$ for $G_{0,n} \in \mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n})$:

$$\begin{aligned} P_{X^n, Y^n}(G_{0,n}) &\triangleq (P_{X^n} \otimes \vec{P}_{Y^n|X^n})(G_{0,n}) \\ &= \int \vec{P}_{Y^n|X^n}(G_{0,n, x^n}|x^n) \otimes P_{X^n}(dx^n) \end{aligned}$$

where G_{0,n, x^n} is the x^n -section of $G_{0,n}$ at point x^n defined by $G_{0,n, x^n} \triangleq \{y^n \in \mathcal{Y}_{0,n} : (x^n, y^n) \in G_{0,n}\}$ and \otimes denotes the convolution.

(P3): The marginal distribution $P_{Y^n} \in \mathcal{M}_1(\mathcal{Y}_{0,n})$:

$$\begin{aligned} P_{Y^n}(F_{0,n}) &\triangleq P(\mathcal{X}_{0,n} \times F_{0,n}), \quad F_{0,n} \in \mathcal{B}(\mathcal{Y}_{0,n}) \\ &= \int \vec{P}_{Y^n|X^n}((\mathcal{X}_{0,n} \times F_{0,n})_{x^n}; x^n) P_{X^n}(dx^n) \\ &= \int \vec{P}_{Y^n|X^n}(F_{0,n}|x^n) P_{X^n}(dx^n). \end{aligned}$$

The set of all $(n+1)$ -fold such convolution distributions is defined by

$$\begin{aligned} \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) &= \left\{ \vec{P}_{Y^n|X^n}(dy^n|x^n) \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) : \right. \\ &\left. \vec{P}_{Y^n|X^n}(dy^n|x^n) \triangleq \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) - a.s. \right\}. \end{aligned}$$

Directed information is defined via the Kullback-Leibler distance:

$$\begin{aligned} I(X^n \rightarrow Y^n) &\triangleq \mathbb{D}(P_{X^n, Y^n} || P_{X^n} \times P_{Y^n}) \\ &= \mathbb{D}(P_{X^n} \otimes \vec{P}_{Y^n|X^n} || P_{X^n} \times P_{Y^n}) \\ &= \int \log \left(\frac{d(P_{X^n} \otimes \vec{P}_{Y^n|X^n})}{d(P_{X^n} \times P_{Y^n})} \right) d(P_{X^n} \otimes \vec{P}_{Y^n|X^n}) \\ &= \int \log \left(\frac{\vec{P}_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)} \right) \vec{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \\ &\equiv \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n}). \end{aligned} \quad (7)$$

Note that (7) states that directed information is expressed as a functional of $\{P_{X^n}, \vec{P}_{Y^n|X^n}\}$.

Next, the definition of nonanticipative RDF is given.

Definition 1: (Nonanticipative Rate Distortion Function) Suppose $d_{0,n} \triangleq \sum_{i=0}^n \rho_{0,i}(x^i, y^i)$, where $\rho_{0,i} : \mathcal{X}_{0,i} \times \mathcal{Y}_{0,i} \rightarrow [0, \infty)$, is a sequence of $\mathcal{B}(\mathcal{X}_{0,i}) \times \mathcal{B}(\mathcal{Y}_{0,i})$ -measurable distortion functions, and

let $\vec{\mathcal{Q}}_{0,n}(D)$ (assuming is non-empty) denotes the average distortion or fidelity constraint defined by

$$\begin{aligned} \vec{\mathcal{Q}}_{0,n}(D) &\triangleq \left\{ \vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) : \ell_{d_{0,n}}(\vec{P}_{Y^n|X^n}) \right. \\ &\triangleq \left. \int d_{0,n}(x^n, y^n) \vec{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \leq D \right\} \end{aligned} \quad (8)$$

where $D \geq 0$. The nonanticipative RDF is defined by

$$R_{0,n}^c(D) \triangleq \inf_{\vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}_{0,n}(D)} \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n}). \quad (9)$$

Clearly, $R_{0,n}^c(D)$ is characterized by minimizing directed information or equivalently $\mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n})$ over $\vec{\mathcal{Q}}_{0,n}(D)$.

III. EXISTENCE OF RECONSTRUCTION CONDITIONAL DISTRIBUTION

In this section, the existence of the minimizing $(n+1)$ -fold convolution of conditional distributions in (9) is established by using the topology of weak convergence of probability measures on Polish spaces. Before we present the relevant results we state some properties of average distortion set $\vec{\mathcal{Q}}_{0,n}(D)$ and directed information $\mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n})$. These properties are derived in [14].

Theorem 1: [14] Let $\{\mathcal{X}_n : n \in \mathbb{N}\}$ and $\{\mathcal{Y}_n : n \in \mathbb{N}\}$ be Polish spaces. Then

- (1) The set $\vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ is convex.
- (2) $\mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n})$ is a convex functional of $\vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ for a fixed $P_{X^n} \in \mathcal{M}_1(\mathcal{X}_{0,n})$.
- (3) The set $\vec{\mathcal{Q}}_{0,n}(D)$ is convex.

Let $BC(\mathcal{Y}_{0,n})$ denotes the set of bounded continuous real-valued functions on $\mathcal{Y}_{0,n}$. Below, we introduce the main conditions for establishing existence of nonanticipative RDF (9).

Assumption 1: The following conditions are assumed throughout the paper.

- (1) $\mathcal{Y}_{0,n}$ is a compact Polish space, $\mathcal{X}_{0,n}$ is a Polish space;
- (2) for all $h(\cdot) \in BC(\mathcal{Y}_{0,n})$, the function mapping $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1} \mapsto \int_{\mathcal{Y}_n} h(y) P_{Y|Y^{n-1}, X^n}(dy|y^{n-1}, x^n) \in \mathbb{R}$ is continuous jointly in the variables $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1}$;
- (3) $d_{0,n}(x^n, \cdot)$ is continuous on $\mathcal{Y}_{0,n}$;
- (4) the distortion level D is such that there exist sequence $(x^n, y^n) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}$ satisfying $d_{0,n}(x^n, y^n) < D$.

Note that since $\mathcal{Y}_{0,n}$ is assumed to be a compact Polish space, then by [13] probability measures on $\mathcal{Y}_{0,n}$ are weakly compact. Moreover, the following weak compactness result can be obtained, which will be used to show existence of an optimal nonanticipative RDF, $R_{0,n}^c(D)$.

Lemma 1: Suppose Assumption 1 (1), (2) hold.

Then

- (1) The set $\vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ is weakly compact.
- (2) Under the additional conditions (3), (4) the set $\vec{\mathcal{Q}}_{0,n}(D)$ is a closed subset of $\vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ (hence compact).

Proof: (1) This follows from the fact that any $\vec{P}_{Y^n|X^n}(dy^n|x^n) \in \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ is factorized as $\vec{P}_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i)$ -a.s., where $P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i})$, $1 \leq i \leq n$, and $\mathcal{Y}_{0,n}$ compact Polish space implies that $\{P_{Y_i|Y^{i-1}, X^i}(\cdot|y^{i-1}, x^i) : y^{i-1} \in \mathcal{Y}_{0,i-1}, x^i \in \mathcal{X}_{0,i}\}$ is compact, hence by Prohorov's theorem it is uniformly tight $\forall i$. Utilizing this, by induction it can be shown that the family of convolution measures $\vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ is compact.

(2) Utilizing compactness of $\vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ and condition (3) of Assumption 1 on $d_{0,n}(x^n, \cdot)$, it can be shown that $\vec{\mathcal{Q}}_{0,n}(D)$ is a closed subset of $\vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$. ■

The previous results follow from Prohorov's theorem that relates tightness and weak compactness.

The next theorem establishes existence of the minimizing reconstruction kernel for (9); it follows from Lemma 1 and the lower semicontinuity of $\mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \cdot)$ with respect to $\vec{P}_{Y^n|X^n}$.

Theorem 2: Suppose the conditions of Lemma 1 hold. Then $R_{0,n}^c(D)$ has a minimum.

Proof: The proof is omitted due to space limitations. ■

IV. OPTIMAL RECONSTRUCTION OF NONANTICIPATIVE RATE DISTORTION FUNCTION

In this section the form of the optimal reconstruction conditional distribution is derived under a stationarity assumption. The method is based on calculus of variations on the space of measures. We introduce the following main assumption.

Assumption 2: The $(n+1)$ -fold convolution conditional distribution $\vec{P}_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i)$ -a.s., is the convolution of stationary conditional distributions.

Assumption 2 holds for stationary process $\{(X_i, Y_i) : i \in \mathbb{N}\}$ and $\rho_{0,i}(x^i, y^i) \equiv \rho(T^i x^n, T^i y^n)$, where $T^i x^n$ is the shift operator on x^n (and similarly for $T^i y^n$). The consequence of Assumption 2, which holds for stationary processes and a single letter distortion function, is that the Gateaux differential of $\mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n})$ is done in only one direction (since $P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i)$ are stationary). Therefore, we define the variation of $\vec{P}_{Y^n|X^n}$ in the direction of $\vec{P}_{Y^n|X^n} - \vec{P}_{Y^n|X^n}^0$ via $\vec{P}_{Y^n|X^n}^\epsilon \triangleq \vec{P}_{Y^n|X^n} + \epsilon(\vec{P}_{Y^n|X^n} - \vec{P}_{Y^n|X^n}^0)$, $\epsilon \in [0, 1]$, since under Assumption 2, the functionals $\{P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i) \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i}) : i = 0, 1, \dots, n\}$ are identical.

Theorem 3: Suppose Assumption 2 holds and $\mathbb{I}_{P_{X^n}}(\vec{P}_{Y^n|X^n}) \triangleq \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n})$ is well defined for every $\vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}_{0,n}(D)$

possibly taking values from the set $[0, \infty)$. Then $\vec{P}_{Y^n|X^n} \rightarrow \mathbb{I}_{P_{X^n}}(\vec{P}_{Y^n|X^n})$ is Gateaux differentiable at every point in $\vec{\mathcal{Q}}_{0,n}(D)$, and the Gateaux derivative at the point $\vec{P}_{Y^n|X^n}^0$ in the direction $\vec{P}_{Y^n|X^n} - \vec{P}_{Y^n|X^n}^0$ is given by

$$\begin{aligned} & \delta \mathbb{I}_{P_{X^n}}(\vec{P}_{Y^n|X^n}^0, \vec{P}_{Y^n|X^n} - \vec{P}_{Y^n|X^n}^0) \\ &= \int \log \left(\frac{\vec{P}_{Y^n|X^n}^0(dy^n|x^n)}{P_{Y^n}^0(dy^n)} \right) \\ & \quad \otimes (\vec{P}_{Y^n|X^n} - \vec{P}_{Y^n|X^n}^0)(dy^n|x^n) P_{X^n}(dx^n) \end{aligned}$$

where $P_{Y^n}^0 \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is the marginal measure corresponding to $\vec{P}_{Y^n|X^n}^0 \otimes P_{X^n} \in \mathcal{M}_1(\mathcal{Y}_{0,n} \times \mathcal{X}_{0,n})$.

Proof: The proof is similar to the one in [15] (although it is more involved). ■

The constrained problem defined by (9) can be reformulated as an unconstrained problem using Lagrange multipliers. The equivalence of constrained and unconstrained problems is established next.

Lemma 2: Suppose Assumptions 1, 2 hold and consider $d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho(T^i x^n, T^i y^n)$, where $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \rightarrow \bar{R}_0 \equiv [0, \infty]$ is continuous in the second argument. Then the constrained problem as stated in Theorem 2, is equivalent to an unconstrained problem stated below.

$$\begin{aligned} & \inf_{\vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}_{0,n}(D)} \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n}) \\ &= \max_{s \leq 0} \inf_{\vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})} \left\{ \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n}) \right. \\ & \quad \left. - s \ell_{d_{0,n}}(\vec{P}_{Y^n|X^n}) \right\} \\ &= \max_{s \leq 0} \inf_{\vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})} \left\{ \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n}) \right. \\ & \quad \left. - s \left(\int d_{0,n}(x^n, y^n) \vec{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) - D \right) \right\}. \end{aligned}$$

Further the infimum occurs on the boundary of the set $\vec{\mathcal{Q}}_{0,n}(D)$.

Proof: The proof utilizes the Lagrange duality theorem [16]. ■

Utilizing Lemma 2, then

$$\begin{aligned} R_{0,n}^c(D) &= \sup_{s \leq 0} \inf_{\substack{\vec{P}_{Y^n|X^n} \\ \in \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})}} \left\{ \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \vec{P}_{Y^n|X^n}) \right. \\ & \quad \left. - s(\ell_{d_{0,n}}(\vec{P}_{Y^n|X^n}) - D) \right\}. \end{aligned} \quad (10)$$

Note that $\vec{P}_{Y^n|X^n} \in \vec{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ are probability measures on $\mathcal{Y}_{0,n}$ therefore, one should introduce another set of Lagrange multipliers to obtain an unconstrained problem free of such a constraint.

Since $\vec{P}_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}(dy_i|y^{i-1}, x^i)$ is a consistent probability measure on $\mathcal{Y}_{0,n}$, then for each $k = 0, 1, \dots, n$, $\int_{\mathcal{Y}_{0,k}} \vec{P}_{Y^k|X^k}(dy^k|x^k) = 1$. This constraint

is expressed via

$$\begin{aligned} & \sum_{i=0}^n \int \lambda_i(x^i, y^{i-1}) \left(\vec{P}_{Y^i|X^i}(dy^i|x^i) - 1 \right) P_{X^i}(dx^i) \\ &= \sum_{i=0}^n \int \lambda_i(x^i, y^{i-1}) \left(\vec{P}_{Y^n|X^n}(dy^n|x^n) - 1 \right) P_{X^n}(dx^n) \end{aligned} \quad (11)$$

where $\{\lambda_i(\cdot, \cdot) : i = 0, 1, \dots, n\}$ are Lagrange multipliers. The above observations yield the following theorem.

Theorem 4: Suppose the Assumptions of Lemma 2 hold and consider $d_{0,n}(x^n, y^n) \triangleq \sum_{i=0}^n \rho(T^i x^n, T^i y^n)$. Then

- (1) The infimum in (10) is attained at $\vec{P}_{Y^n|X^n}^* \in \vec{Q}_{0,n}(D)$ given by⁵

$$\begin{aligned} \vec{P}_{Y^n|X^n}^*(dy^n|x^n) &= \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}^*(dy_i|y^{i-1}, x^i) \\ &= \otimes_{i=0}^n \frac{e^{s\rho(T^i x^n, T^i y^n)} P_{Y_i|Y^{i-1}}^*(dy_i|y^{i-1})}{\int_{\mathcal{Y}_i} e^{s\rho(T^i x^n, T^i y^n)} P_{Y_i|Y^{i-1}}^*(dy_i|y^{i-1})} \end{aligned} \quad (12)$$

where $s \leq 0$ and $P_{Y_i|Y^{i-1}}^*(dy_i|y^{i-1}) \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1})$.

- (2) The nonanticipative RDF is given by

$$\begin{aligned} R_{0,n}^c(D) &= sD - \sum_{i=0}^n \int \log \left(\int_{\mathcal{Y}_i} e^{s\rho(T^i x^n, T^i y^n)} \right. \\ &\quad \left. P_{Y_i|Y^{i-1}}^*(dy_i|y^{i-1}) \right) \vec{P}_{Y^{i-1}|X^{i-1}}^*(dy^{i-1}|x^{i-1}) \otimes P_{X^i}(dx^i). \end{aligned} \quad (13)$$

If $R_{0,n}^c(D) > 0$ then $s < 0$ and

$$\sum_{i=0}^n \int \rho(T^i x^n, T^i y^n) \vec{P}_{Y^i|X^i}^*(dy^i; x^i) P_{X^i}(dx^i) = D. \quad (14)$$

Proof: The fully unconstrained problem of (10) is obtained by introducing another set of Lagrange multipliers $\{\lambda_i(\cdot, \cdot) : i = 0, 1, \dots, n\}$ as in (11). The derivation is omitted due to space limitations. ■

Remark 1: Note that if the distortion function satisfies $\rho(T^i x^n, T^i y^n) = \rho(x_i, T^i y^n)$ then for $i = 0, 1, \dots, n$

$$P_{Y_i|Y^{i-1}, X^i}^*(dy_i|y^{i-1}, x^i) = P_{Y_i|Y^{i-1}, X^i}^*(dy_i|y^{i-1}, x_i) - a.s. \quad (15)$$

that is, the reconstruction kernel is Markov in X^n . However, without further restriction one cannot claim that this conditional distribution is also Markov with respect to $\{\mathcal{Y}_i : i = 0, 1, \dots, n\}$.

V. REALIZATION OF NONANTICIPATIVE RATE DISTORTION FUNCTION

The realization of the nonanticipative RDF (optimal reconstruction conditional distribution) is equivalent to the sensor mapping as shown in Fig. 2 which produces the auxiliary random process $\{Z_i : i \in \mathbb{N}\}$ that will be used for filtering. This is equivalent to identifying a communication channel, an encoder and a decoder such that the reconstruction from

⁵Due to stationarity assumption $P_{Y_i|Y^{i-1}}(\cdot|\cdot) = P(\cdot|\cdot)$ and $P_{Y_i|Y^{i-1}, X^i}^*(\cdot|\cdot, \cdot) = P^*(\cdot|\cdot, \cdot)$

the sequence X^n to the sequence Y^n matches the nonanticipative rate distortion minimizing reconstruction kernel. Fig. 3 illustrates the cascade sub-systems that realize the nonanticipative RDF, which is consistent with the discussion in the introduction.

Definition 2: Given a source $\{P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1}) : i = 0, \dots, n\}$, a channel $\{P_{B_i|B^{i-1}, A^i}(db_i|b^{i-1}, a^i) : i = 0, \dots, n\}$ is a realization of the optimal reconstruction distribution if there exists a pre-channel encoder $\{P_{A_i|A^{i-1}, B^{i-1}, X^i}(da_i|a^{i-1}, b^{i-1}, x^i) : i = 0, \dots, n\}$ and a post-channel decoder $\{P_{Y_i|Y^{i-1}, B^i}(dy_i|y^{i-1}, b^i) : i = 0, \dots, n\}$ such that

$$\vec{P}_{Y^n|X^n}^*(dy^n|x^n) \triangleq \otimes_{i=0}^n P_{Y_i|Y^{i-1}, X^i}^*(dy_i|y^{i-1}, x^i) - a.s.$$

where the joint distribution is

$$\begin{aligned} P_{X^n, A^n, B^n, Y^n}(dx^n, da^n, db^n, dy^n) &= \otimes_{i=0}^n P_{Y_i|Y^{i-1}, B^i}(dy_i|y^{i-1}, b^i) \otimes P_{B_i|B^{i-1}, A^i}(db_i|b^{i-1}, a^i) \\ &\quad \otimes P_{A_i|A^{i-1}, B^{i-1}, X^i}(da_i|a^{i-1}, b^{i-1}, x^i) \\ &\quad \otimes P_{X_i|X^{i-1}, Y^{i-1}}(dx_i|x^{i-1}, y^{i-1}) - a.s. \end{aligned}$$

The filter is given by $\{P_{X_i|B^{i-1}}(dx_i|b^{i-1}) : i = 0, \dots, n\}$ or by $\{P_{X_i|Y^{i-1}}(dx_i|y^{i-1}) : i = 0, \dots, n\}$.

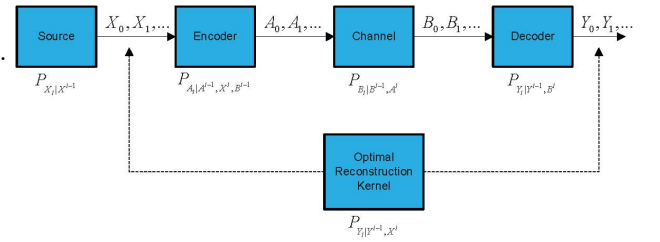


Fig. 3. Realizable nonanticipative rate distortion function.

Thus, if $\{P_{B_i|B^{i-1}, A^i}(db_i|b^{i-1}, a^i) : i = 0, \dots, n\}$ is a realization of the nonanticipative RDF minimizing distribution then the channel connecting the source, encoder, channel, decoder achieves the nonanticipative RDF, and the filter is obtained. Clearly, $\{B_i : i = 0, 1, \dots, n\}$ is an auxiliary random process which is needed to obtain the filter $\{P_{X_i|B^{i-1}}(dx_i|b^{i-1}) : i = 0, \dots, n\}$.

In the next section, we provide an example for such a realization.

VI. EXAMPLE

Consider the following discrete-time partially observed linear Gauss-Markov system described by

$$\begin{cases} X_{t+1} = AX_t + BW_t, & X_0 = X, \quad t \in \mathbb{N}^n \\ Y_t = CX_t + DV_t, & t \in \mathbb{N}^n \end{cases} \quad (16)$$

where $X_t \in \mathbb{R}^m$ is the state (unobserved) process of information source (plant), and $Y_t \in \mathbb{R}^p$ is the partially observed (measurement) process. Assume that (C, A) is detectable and $(A, BB^{tr})^{\frac{1}{2}}$ is stabilizable, $(D \neq 0)$. The state and observation noise $\{(W_t, V_t) : t \in \mathbb{N}^n\}$ are mutually independent, independent of the Gaussian RV X_0 , with parameters $N(\bar{x}_0, \bar{V}_0)$, where $W_t \in \mathbb{R}^k$ and $V_t \in \mathbb{R}^d$,

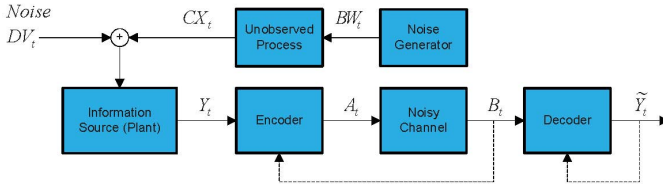


Fig. 4. Communication system.

are Gaussian IID processes with zero mean and identity covariances.

The realization will be done following Fig. 4. The objective is to reconstruct $\{Y_t : t \in \mathbb{N}^n\}$ by $\{\tilde{Y}_t : t \in \mathbb{N}^n\}$ via nonanticipative operations. The distortion is single letter defined by

$$d_{0,n}(y^n, \tilde{y}^n) \triangleq \frac{1}{n+1} \sum_{i=0}^n \|y_i - \tilde{y}_i\|^2.$$

The objective is to compute

$$R_{0,n}^c(D) = \inf_{\vec{P}_{\tilde{Y}^n|Y^n} \in \vec{\mathcal{Q}}_{0,n}(D)} \mathbb{I}_{X^n \rightarrow Y^n}(P_{Y^n}, \vec{P}_{\tilde{Y}^n|Y^n})$$

and then realize the reconstruction distribution. The filter realization procedure is similar to the one found in reconstruction of $\{X_t : t \in \mathbb{N}^n\}$ in [17]. The methodology however, is based on the explicit formulae of optimal reconstruction of Theorem 4. According to Theorem 4, the optimal reconstruction is given by

$$\vec{P}_{\tilde{Y}^n|Y^n}^*(d\tilde{y}^n|y^n) = \otimes_{i=0}^n \frac{e^{s\|\tilde{y}_i - y_i\|^2} P_{\tilde{Y}_i|\tilde{Y}^{i-1}}(d\tilde{y}_i|\tilde{y}^{i-1})}{\int_{\mathcal{Y}_i} e^{s\|\tilde{y}_i - y_i\|^2} P_{\tilde{Y}_i|\tilde{Y}^{i-1}}(d\tilde{y}_i|\tilde{y}^{i-1})} \quad (17)$$

where $s \leq 0$. Hence, from (17) it follows that $P_{\tilde{Y}_i|\tilde{Y}^{i-1}, Y^i} = P_{\tilde{Y}_i|\tilde{Y}^{i-1}, Y_i}(d\tilde{y}_i|\tilde{y}^{i-1}, y_i)$ —a.s., that is the reconstruction is Markov with respect to the process $\{Y_i : i \in \mathbb{N}^n\}$. Moreover, since the exponential term $\|\tilde{y}_i - y_i\|^2$ in the RHS of (17) is quadratic in (y_i, \tilde{y}_i) , and $\{X_i : i \in \mathbb{N}^n\}$ is Gaussian, then $\{(X_i, Y_i) : i \in \mathbb{N}^n\}$ is jointly Gaussian, hence it follows that $P_{\tilde{Y}_i|\tilde{Y}^{i-1}, Y_i}(\cdot|\tilde{y}^{i-1}, y_i)$ is Gaussian (for a fixed realization of (\tilde{y}^{i-1}, y_i)). Hence, it has the general form

$$\tilde{Y}_t = \bar{A}Y_t + \bar{B}\tilde{Y}^{t-1} + Z_t, \quad t \in \mathbb{N}^n \quad (18)$$

where $\bar{A}_t \in \mathbb{R}^{p \times p}$, $\bar{B}_t \in \mathbb{R}^{p \times tp}$, and $\{Z_t : t \in \mathbb{N}^n\}$ is an independent sequence of Gaussian vectors. The channel in (18) can be realized as follows.

The communication channel (18) can be realized via a scalar additive Gaussian noise channel with feedback defined by

$$B_t = A_t + Z_t, \quad t \in \mathbb{N}^n \quad (19)$$

where the encoder is a mapping $A_t = \Phi_t(Y_t, \tilde{Y}^{t-1})$ with power $P_t \triangleq \text{Tr}\{E\{(A_t)^2\}\}$. For A_t Gaussian the directed information is $I(A^t \rightarrow B^t) = \log|1 + E\{(A_t)^2\} \text{Cov}(Z_t)^{-1}|$. The decoder at time $t \in \mathbb{N}^n$ receives B^t and computes the reconstruction $\tilde{Y}_t = \Psi_t(B^t, \tilde{Y}^{t-1})$.

Realization of the nonanticipative RDF. The realization is

based on the block diagram of Fig. 5. The encoder $\Phi_t(\cdot, \cdot)$ consists of a pre-encoder which produces the Gaussian innovation process $\{K_t : t \in \mathbb{N}^n\}$, defined by

$$K_t \triangleq Y_t - E\{Y_t|\sigma\{\tilde{Y}^{t-1}\}\}, \quad t \in \mathbb{N}^n \quad (20)$$

whose covariance is defined by $\Lambda_t \triangleq E\{K_t K_t^{tr}\}$. The decoder consists of a pre-decoder $\{\tilde{K}_t : t \in \mathbb{N}^n\}$ which is defined by

$$\tilde{K}_t \triangleq \tilde{Y}_t - E\{\tilde{Y}_t|\sigma\{\tilde{Y}^{t-1}\}\}, \quad t \in \mathbb{N}. \quad (21)$$

Note that the fidelity criterion satisfies $d_{0,n}(y^n, \tilde{y}^n) = d_{0,n}(k^n, \tilde{k}^n) = \frac{1}{n+1} \sum_{i=0}^n \|\tilde{k}_i - k_i\|^2$. Let $\{E_t : t \in \mathbb{N}\}$ be the unitary matrix that diagonalizes $\{\Lambda_t : t \in \mathbb{N}^n\}$, such that

$$E_t \Lambda_t E_t^{tr} = \text{diag}\{\lambda_{t,1}, \dots, \lambda_{t,p}\}, \quad t \in \mathbb{N}^n. \quad (22)$$

Choose $\{\xi_t : t \in \mathbb{N}^n\}$ such that

$$\delta_{t,i} \triangleq \begin{cases} \xi_t & \text{if } \xi_t \leq \lambda_{t,i} \\ \lambda_{t,i} & \text{if } \xi_t > \lambda_{t,i} \end{cases}, \quad t \in \mathbb{N}^n, \quad i = 1, \dots, p$$

where $\{\xi_t : t \in \mathbb{N}^n\}$ satisfies $\sum_{i=1}^p \delta_{t,i} = D$.

Define $\Gamma_t \triangleq E_t K_t$. Then $\{\Gamma_t : t \in \mathbb{N}^n\}$ is an orthogonal process. Let $\{\tilde{\Gamma}_t : t \in \mathbb{N}^n\}$ denote its reconstruction and define $d_{0,n}(\Gamma^n, \tilde{\Gamma}^n) \triangleq \frac{1}{n+1} \sum_{i=0}^n \|\Gamma_i - \tilde{\Gamma}_i\|^2$. Then by [18],

$$R_{0,n}^{c, \Gamma^n, \tilde{\Gamma}^n}(D) \triangleq \inf_{\vec{P}_{\tilde{\Gamma}^n|\Gamma^n}} \mathbb{I}_{X^n \rightarrow Y^n}(P_{\Gamma^n}, \vec{P}_{\tilde{\Gamma}^n|\Gamma^n}) E\{d_{0,n}(\Gamma^n, \tilde{\Gamma}^n) \leq D\}$$

has a solution

$$P_{\tilde{\Gamma}^n|\Gamma^n}^*(d\tilde{\gamma}^n|\gamma^n) = \otimes_{i=0}^n P_{\Gamma_i|\tilde{\Gamma}_i}^*(d\tilde{\gamma}_i|\gamma_i) - a.s.,$$

where $P_{\Gamma_i|\tilde{\Gamma}_i}^*(\cdot|\cdot) \sim N(\eta_{t,i}\Gamma_t, \eta_{t,i}\delta_{t,i})$, $\eta_{t,i} \triangleq (1 - \frac{\delta_{t,i}}{\lambda_{t,i}})$, $i = 0, 1, \dots, p$, and $R_{0,n}^{c, \Gamma^n, \tilde{\Gamma}^n}(D) = \frac{1}{n+1} \sum_{i=1}^d \log\left(\frac{\lambda_{t,i}}{\delta_{t,i}}\right)$. Thus, the pre-encoder can be further scaled by $\Gamma_t = E_t K_t$, and Γ_t is compressed by $A_t = \mathcal{A}_t \Gamma_t$ and sent through an additive white Gaussian noise (AWGN) channel with feedback, after which the received signal is decompressed by $\tilde{\Gamma}_t = \mathcal{B}_t B_t$ in the pre-decoder. By the knowledge of the channel output at the decoder, the mean square estimator \hat{X}_t is generated at the decoder (and encoder because $\hat{X}_t \triangleq E\{X_t|\sigma\{\tilde{Y}^{t-1}\}\}$). The complete design is illustrated in Fig. 5. Next we pick a

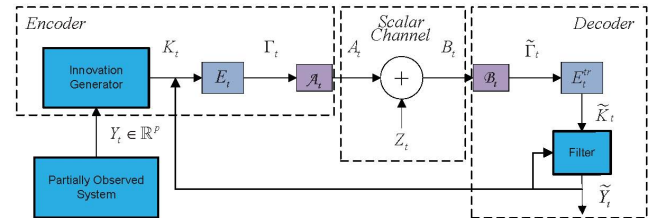


Fig. 5. Design of the discrete-time communication system with scalar additive white Gaussian noise (AWGN) channel.

specific AWGN channel.

Scalar AWGN Channel. Consider a scalar channel $B_t = A_t + Z_t$, $t \in \mathbb{N}^n$, where Z_t is Gaussian zero mean, $Q \triangleq \text{Cov}(Z_t)$, and $A_t \in \mathbb{R}$. We can design $\{(\mathcal{A}_t, \mathcal{B}_t) : t \in \mathbb{N}^n\}$ by

$$\mathcal{A}_t = \left[\sqrt{\frac{\alpha_1 P_t}{\lambda_{t,1}}}, \dots, \sqrt{\frac{\alpha_p P_t}{\lambda_{t,p}}} \right], \quad t \in \mathbb{N}^n$$

$$\mathcal{B}_t = \left[\sqrt{\alpha_1 P_t \lambda_{t,1}}, \dots, \sqrt{\alpha_p P_t \lambda_{t,p}} \right]^{tr}, \quad t \in \mathbb{N}^n$$

where $\sum_{i=1}^p \alpha_i = 1$, $i = 1, \dots, p$. Note that

$$\begin{aligned} H_t &= \mathcal{B}_t \mathcal{A}_t \\ &= \left[\sqrt{\alpha_1 P_t \lambda_{t,1}}, \dots, \sqrt{\alpha_p P_t \lambda_{t,p}} \right]^{tr} \left[\sqrt{\frac{\alpha_1 P_t}{\lambda_{t,1}}}, \dots, \sqrt{\frac{\alpha_p P_t}{\lambda_{t,p}}} \right] \\ &= \left[\begin{array}{c} \sqrt{\alpha_1 P_t \lambda_{t,1}} \\ \dots \\ \sqrt{\alpha_p P_t \lambda_{t,p}} \end{array} \right] \left[\sqrt{\frac{\alpha_1 P_t}{\lambda_{t,1}}}, \dots, \sqrt{\frac{\alpha_p P_t}{\lambda_{t,p}}} \right] \\ &= P_t \left[\begin{array}{ccc} \alpha_1 & \dots & \sqrt{\alpha_1 \alpha_p \frac{\lambda_{t,1}}{\lambda_{t,p}}} \\ \vdots & \dots & \vdots \\ \sqrt{\alpha_p \alpha_1 \frac{\lambda_{t,p}}{\lambda_{t,1}}} & \dots & \alpha_p \end{array} \right] \in \mathbb{R}^{p \times p}. \end{aligned}$$

Therefore,

$$\tilde{\Gamma}_t = H_t E_t K_t + \mathcal{B}_t Z_t, \quad \Gamma_t = E_t K_t, \quad t \in \mathbb{N}^n. \quad (23)$$

By pre-multiplying $\tilde{\Gamma}_t$ by E_t^{tr} we can construct

$$\begin{aligned} \tilde{K}_t &= E_t^{tr} \tilde{\Gamma}_t \\ &= E_t^{tr} H_t E_t K_t + E_t^{tr} \mathcal{B}_t Z_t, \quad t \in \mathbb{N}^n. \end{aligned}$$

The reconstruction of Y_t is given by the sum of \tilde{K}_t and $C \hat{X}_t$ as follows.

$$\begin{aligned} \tilde{Y}_t &= \Psi_t(B^t, \tilde{Y}^{t-1}) \\ &= \tilde{K}_t + C \hat{X}_t, \quad \hat{X}_t = E\{X_t | \sigma\{\tilde{Y}^{t-1}\}\} \\ &= E_t^{tr} H_t E_t K_t + E_t^{tr} \mathcal{B}_t Z_t + C \hat{X}_t, \quad t \in \mathbb{N}^n. \end{aligned} \quad (24)$$

Next, it will be shown that the desired distortion is achieved by the above realization while the filter of $\{Y_t : t \in \mathbb{N}^n\}$ is based on $\{\tilde{Y}_t : t \in \mathbb{N}^n\}$ given by (25).

First, we notice that

$$E\{(Y_t - \tilde{Y}_t)^{tr} (Y_t - \tilde{Y}_t)\} = \text{Tr}\left(E\{(Y_t - \tilde{Y}_t)(Y_t - \tilde{Y}_t)^{tr}\}\right).$$

Then we can compute

$$\begin{aligned} E\{(Y_t - \tilde{Y}_t)^{tr} (Y_t - \tilde{Y}_t)\} &= \text{Tr} E\{(K_t - \tilde{K}_t)(K_t - \tilde{K}_t)^{tr}\} \\ &= \text{Tr} E\{(K_t - E_t^{tr} \tilde{\Gamma}_t)(K_t - E_t^{tr} \tilde{\Gamma}_t)^{tr}\} \\ &= \text{Tr} E\{(K_t - E_t^{tr} H_t E_t K_t - E_t^{tr} \mathcal{B}_t Z_t) \\ &\quad (K_t - E_t^{tr} H_t E_t K_t - E_t^{tr} \mathcal{B}_t Z_t)^{tr}\} \\ &= \text{Tr} E\{(I - E_t^{tr} H_t E_t) K_t - E_t^{tr} \mathcal{B}_t Z_t \\ &\quad ((I - E_t^{tr} H_t E_t) K_t - E_t^{tr} \mathcal{B}_t Z_t)^{tr}\} \\ &= \text{Tr}\left\{(I - E_t^{tr} H_t E_t) \Lambda_t (I - E_t^{tr} H_t E_t)^{tr} \right. \\ &\quad \left. + E_t^{tr} \mathcal{B}_t Q \mathcal{B}_t^{tr} E_t\right\} \\ &= \text{Tr}\left\{(I - E_t^{tr} H_t E_t) E_t^{tr} \text{diag}(\lambda_{t,1}, \dots, \lambda_{t,p}) \right. \\ &\quad \left. E_t (I - E_t^{tr} H_t E_t)^{tr} + E_t^{tr} \mathcal{B}_t Q \mathcal{B}_t^{tr} E_t\right\} \\ &= \text{Tr}\left\{E_t^{tr} \left((I - H_t) \text{diag}(\lambda_{t,1}, \dots, \lambda_{t,p}) (1 - H_t)^{tr} \right. \right. \\ &\quad \left. \left. + (\mathcal{B}_t Q \mathcal{B}_t^{tr})\right) E_t\right\} \\ &= \text{Tr}\left\{\text{diag}(\delta_{t,1}, \dots, \delta_{t,p})\right\} = D. \end{aligned}$$

Decoder. The decoder is $\tilde{Y}_t = \tilde{K}_t + C \hat{X}_t$, where $\hat{X}_t : t \in \mathbb{N}^n$ is obtained from the modified Kalman filter as follows. Recall that

$$\begin{aligned} \tilde{Y}_t &= \tilde{K}_t + C \hat{X}_t \\ &= E_t^{tr} H_t E_t (Y_t - C \hat{X}_t) + E_t^{tr} \mathcal{B}_t Z_t + C \hat{X}_t \\ &= E_t^{tr} H_t E_t (C X_t + D V_t - C \hat{X}_t) + E_t^{tr} \mathcal{B}_t Z_t + C \hat{X}_t \\ &= E_t^{tr} H_t E_t C X_t - E_t^{tr} H_t E_t C \hat{X}_t + C \hat{X}_t \\ &\quad + (E_t^{tr} H_t E_t D V_t + E_t^{tr} \mathcal{B}_t Z_t) \end{aligned}$$

where $\{V_t : t \in \mathbb{N}^n\}$ and $\{Z_t : t \in \mathbb{N}^n\}$ are independent Gaussian vectors. Then $\hat{X}_t = E\{X_t | \sigma\{\tilde{Y}^{t-1}\}\}$ is given by the modified Kalman filter

$$\begin{aligned} \hat{X}_{t+1} &= A \hat{X}_t + C \hat{X}_t + A \Sigma_t (E_t^{tr} H_t E_t C)^{tr} M_t^{-1} \tilde{Y}_t, \quad \hat{X}_0 = \bar{x}_0 \\ \Sigma_{t+1} &= A \Sigma_t A^{tr} - A \Sigma_t (E_t^{tr} H_t E_t C)^{tr} M_t^{-1} (E_t^{tr} H_t E_t C) \Sigma_t A \\ &\quad + B B_t^{tr}, \quad \Sigma_0 = \bar{\Sigma}_0 \end{aligned}$$

where

$$\begin{aligned} M_t &= E_t^{tr} H_t E_t C \Sigma_t (E_t^{tr} H_t E_t C)^{tr} \\ &\quad + E_t^{tr} H_t E_t D D^{tr} (E_t^{tr} H_t E_t)^{tr} + E_t^{tr} \mathcal{B}_t \Sigma_t \mathcal{B}_t^{tr} E_t^{tr}. \end{aligned}$$

Infinite Horizon. As $t \rightarrow \infty$, under the assumption that the linear Gauss-Markov system is stabilizable and detectable, we have

$$\begin{aligned} \Sigma_\infty &= A \Sigma_\infty A^{tr} \\ &\quad - A \Sigma_\infty (E_\infty^{tr} H_\infty E_\infty C)^{tr} M_\infty^{-1} (E_\infty^{tr} H_\infty E_\infty C) \Sigma_\infty A \\ &\quad + B B_\infty^{tr} \end{aligned}$$

where

$$\begin{aligned} M_\infty &= E_\infty^{tr} H_\infty E_\infty C \Sigma_\infty (E_\infty^{tr} H_\infty E_\infty C)^{tr} \\ &\quad + E_\infty^{tr} H_\infty E_\infty D D^{tr} (E_\infty^{tr} H_\infty E_\infty)^{tr} + E_\infty^{tr} \mathcal{B}_\infty \Sigma_\infty \mathcal{B}_\infty^{tr} E_\infty^{tr} \end{aligned}$$

and E_∞ is the unitary matrix that diagonalizes Λ_∞ by

$$E_\infty \Lambda_\infty E_\infty^{tr} = \text{diag}(\lambda_{\infty,1}, \dots, \lambda_{\infty,p})$$

and

$$\delta_{\infty,i} \triangleq \begin{cases} \xi_\infty & \text{if } \xi_\infty \leq \lambda_{\infty,i} \\ \lambda_{\infty,i} & \text{if } \xi_\infty > \lambda_{\infty,i} \end{cases}, \quad i = 1, \dots, p$$

satisfying $\sum_{i=1}^p \delta_{\infty,i} = D$.

Define

$$\Delta_\infty = \text{diag}(\delta_{\infty,1}, \dots, \delta_{\infty,p}), \quad H_\infty = \text{diag}(\eta_{\infty,1}, \dots, \eta_{\infty,p})$$

where $\eta_{\infty,i} = 1 - \frac{\delta_{\infty,i}}{\lambda_{\infty,i}}$. The realizable (nonanticipative) RDF can be computed as follows.

$$\begin{aligned} R^c(D) &= \lim_{t \rightarrow \infty} \inf_{\substack{P_{\tilde{Y}^t|Y^t}(dy^t|\tilde{y}^t) \\ \in \tilde{\mathcal{Q}}_{0,t}(D)}} \frac{1}{t+1} \mathbb{I}_{X^n \rightarrow Y^n}(P_{Y^t}, \vec{P}_{\tilde{Y}^t|Y^t}) \\ &= \lim_{t \rightarrow \infty} \left(\frac{1}{2} \frac{1}{t+1} \sum_{i=1}^p \log \left(\frac{\lambda_{t,i}}{\delta_{t,i}} \right) \right) \\ &= \frac{1}{2} \sum_{i=1}^p \log \left(\frac{\lambda_{\infty,i}}{\delta_{\infty,i}} \right) \\ &= \frac{1}{2} \log \frac{|\Lambda_\infty|}{|\Delta_\infty|}. \end{aligned}$$

The power constraint satisfies $\text{Tr}\{E\{(A_t)^2\}\} = P_t$, $\lim_{t \rightarrow \infty} P_t = P$. Since $A_t = \mathcal{A}_t E_t K_t$ the capacity is

$$\begin{aligned} C &= \lim_{t \rightarrow \infty} \frac{1}{t+1} I(A^t \rightarrow B^t) \\ &= \frac{1}{2} \log \lim_{t \rightarrow \infty} \frac{1}{t+1} |1 + E\{(A_t)^2\} Q^{-1}| \\ &= \frac{1}{2} \log \lim_{t \rightarrow \infty} \frac{1}{t+1} |1 + E\{(A_t)^2\} Q^{-1}| \\ &= \frac{1}{2} \log \frac{|\Lambda_\infty|}{|\Delta_\infty|} = R^c(D). \end{aligned} \quad (26)$$

Thus, for a given distortion level D , $C = R^c(D)$ is the minimum capacity under which there exists a realizable filter for the data reconstruction of $\{Y_t : t \in \mathbb{N}\}$ by $\{\tilde{Y}_t : t \in \mathbb{N}\}$ ensuring an average distortion equal to D . The filter of $\{X_i : i \in \mathbb{N}\}$ or $\{Y_i : i \in \mathbb{N}\}$ is obtained for $\{\tilde{Y}_i : i \in \mathbb{N}\}$ given by (23) or the auxiliary data $B_i = A_i(Y_i, \tilde{Y}^{i-1}) + Z_i$, $i \in \mathbb{N}$.

VII. CONCLUSION

In this paper, the solution of the nonanticipative RDF is obtained on abstract spaces using the topology of weak convergence of probability measures. A specific example that realizes the optimal causal filter is discussed.

REFERENCES

- [1] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*. Springer-Verlag, Berlin, Heidelberg, New York, 1995.
- [2] R. S. Bucy, "Distortion rate theory and filtering," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 336–340, Mar. 1982.
- [3] A. K. Gorbunov and M. S. Pinsker, "Asymptotic behavior of nonanticipative epsilon-entropy for Gaussian processes," *Problems of Information Transmission*, vol. 27, no. 4, pp. 361–365, 1991.

- [4] J. L. Massey, "Causality, feedback and directed information," in *International Symposium on Information Theory and its Applications (ISITA '90)*, Nov. 27–30 1990, pp. 303–305.
- [5] V. Gupta, A. F. Dana, J. P. Hespanha, R. M. Murray, and B. Hassibi, "Data transmission over networks for estimation and control," *IEEE Transactions on Automatic Control*, vol. 54, no. 8, pp. 1807–1819, Aug. 2009.
- [6] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1056–1068, July 2004.
- [7] G. N. Nair and R. J. Evans, "Stabilizability of Stochastic Linear Systems with Finite Feedback Data Rates," *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 413–436, 2004.
- [8] J. I. Galdos and D. E. Gustafson, "Information and distortion in reduced-order filter design," *IEEE Transactions on Information Theory*, vol. 23, no. 2, pp. 183–194, Mar. 1977.
- [9] X. Feng, A. Loparo, and Y. Fang, "Optimal state estimation for stochastic systems: An information theoretic approach," *IEEE Transactions on Automatic Control*, vol. 42, no. 6, pp. 771–785, June 1997.
- [10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.
- [11] F. H. Lin, K. Hara, V. Solo, M. Vangel, J. W. Belliveau, S. T. Stufflebeam, and H. M. S., "Dynamic Granger-Geweke causality modeling with application to interictal spike propagation," *Human Brain Mapping*, vol. 30, no. 6, pp. 1877–1886, June 2009.
- [12] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 17–44, Feb. 2011.
- [13] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, Inc., New York, 1997.
- [14] C. D. Charalambous and P. A. Stavrou, "Directed information on abstract spaces: properties and extremum problems," in *IEEE International Symposium on Information Theory (ISIT)*, July 1–6 2012, pp. 518–522.
- [15] F. Rezaei, N. U. Ahmed, and C. D. Charalambous, "Rate distortion theory for general sources with potential application to image processing," *International Journal of Applied Mathematical Sciences*, vol. 3, no. 2, pp. 141–165, 2006.
- [16] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, 1969.
- [17] C. D. Charalambous, A. Farhadi, and S. Z. Denic, "Control of continuous-time linear Gaussian systems over additive gaussian wireless fading channels: A separation principle," *IEEE Transactions on Automatic Control*, vol. 53, no. 4, pp. 1013–1019, May 2008.
- [18] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.